MULTISERVER QUEUEING SYSTEMS WITH RETRIALS AND LOSSES

VYACHESLAV M. ABRAMOV

ABSTRACT. The interest to retrial queueing systems is due to their application to telephone systems. The paper studies multiserver retrial queueing systems with n servers. Arrival process is a quite general point process. An arriving customer occupies one of free servers. If upon arrival all servers are busy, then the customer waits for his service in orbit, and after random time retries more and more to occupy a server. The orbit has one waiting space only, and arriving customer, who finds all servers busy and the waiting space occupied, losses from the system. Time intervals between possible retrials are assumed to have arbitrary distribution (the retrial scheme is exactly explained in the paper). The paper provides analysis of this system. Specifically the paper studies optimal number of servers to decrease the loss proportion to a given value. The representation obtained for loss proportion enables us to solve the problem numerically. The algorithm for numerical solution includes effective simulation, which meets the challenge of rare events problem in simulation.

CONTENTS

1. Introduction

The paper studies multiserver queueing systems with retrials and losses. Retrial queueing systems are an object of investigation in numerous papers (e.g. see reviews Artalejo [2], Artalejo and Falin [3], Falin [10]). The interest to these systems is due to their application to telephone systems.

During the last years there has been an especial interest to queueing systems with retrial and losses. The different queueing systems of this type has been studied in [6], [13], [17], and other papers.

The most of known studies in this direction are devoted to analytical solution of one or other problem. Bocharov et al [6] studied multiclass single-server M/G/1 queueing system with finite buffer and limited number of places in orbit. They deduced relationship between steady state distributions of this system and the similar system with only one class of customers. Kovalenko [13] studied the loss probability in M/G/n retrials queueing system with two customer classes. Mandelbaum et al [17] studied queueing system with retrials and losses with time-dependent parameters in framework of Markovian service network (see [16]. They provide fluid and diffusion approximations to obtain numerical characteristics for queue-length and virtual waiting time processes. Atencia and Phong [4] and Bocharov, Phong and Atencia [7] also studied characteristics of queueing systems with retrial and losses with the aid of analytic methods.

However explicit representations can be deduced for only restricted classes of retrial queueing systems, and application of these results for real systems is very difficult.

Recently Abramov [1] studied the main multiserver retrial queueing system, where input stream was a quite general point process (see [1] for more details). The martingale approach, that has been used there, can be adapted to systems with retrials and losses as well, and the present paper is just devoted to analysis of these systems. The exact description of the system is as follows.

• The arrival process $A(t)$ is a point process (the assumptions on this process are given later).

• There are n servers, and an arriving customer occupies one of free servers.

• If upon arrival all servers are busy, but the secondary queue, orbit, having only one space is free, then the customer occupies the orbit and retries more and more to occupy a server.

• A customer, who upon arrival finds all servers busy and the orbit occupied, loses from the system without delay.

• A service time of each customer is exponentially distributed random variable with parameter μ.

• A time between retrials is arbitrary distributed. It is generated by point process $D(t)$ (the details on this assumption are given later). The processes $D(t)$ and $A(t)$ are assumed to be disjoint (that is the probability of simultaneous jump of these processes is equal to 0).

Notice, the significant difference between the main multiserver model of [1] and the model considered in the present paper is also that retrial times in the present model are generally distributed, while in [1] retrial times of each customer in orbit are exponentially distributed.

There is also difference in the assumptions. In the case of the main multiserver retrial queueing systems considered in [1] the point process $A(t)$ must satisfy the condition

$$(1.1) \qquad \mathbb{P}\left\{\lim_{t\to\infty}\frac{A(t)}{t}=\lambda<n\mu\right\}=1,$$

leading to the stability of the system.

In the case of the system considered in the present paper the number of customers in the systems is always bounded, and the stability condition (1.1) is not longer required, and the following more weaker conditions of convergence are used

$$(1.2) \qquad \mathbb{P}\left\{\lim_{t\to\infty}\frac{A(t)}{t}=\lambda\right\}=1,$$

and

$$(1.3) \qquad \mathbb{P}\left\{\lim_{t\to\infty}\frac{D(t)}{t}=\Delta\right\}=1,$$

i.e. λ, Δ, μ and n are not allied to one another.

Conditions (1.2) and (1.3) are technical conditions, that are then used by the methods of our analysis. (We often use the Lebesgue theorem on dominated convergence.)

The main aim of this paper is the numerical solution of a concrete problem associated with specific performance measure. The main question that we want to answer in this paper is *under what number of servers the loss proportion will be less than a given small value?* The system that will be studied in this paper is described by quite general point processes, and explicit analytic solution for this system in general case is impossible. At the same time the direct simulation of this system in order to answer the aforementioned question is not available either, since as the loss probability is chosen very small, we should deal with simulating rare event problem. The simulating rare events problem is a well-known problem having notable attention in the literature (e.g. [12], [23] and many others). Then the main result of this paper is an answer the question *how to simulate this system?*

The main results of the present paper are represented by relation (3.8) and Theorem 4.1. These results enable us to *effectively simulate* models to study their most significant performance characteristic, the loss proportion. By effective simulation we mean such an approach that make the challenge of the rare events problem in simulation.

In our case, we discuss three relations for loss proportions: (3.8), (3.9) and (3.10). Relations (3.9) and (3.10) are not effective. For example, (3.10) requires to simulate the events $\{Q_1(s-) = n\}$ and $\{Q_2(s-) = 1\}$ directly. These events become rare as n increases indefinitely. In contrast relation (3.8) is effective. Its expression contains $\lim_{t\to\infty} t^{-1}\mathbb{E}\int_0^t Q(s)\mathrm{d}s$, which according to Theorem 4.1 can be easily approximated, so that the probabilities of rare events presented in calculation of this expression can be removed. Specifically, there can be used a change of the Poisson measure in (3.2) for calculation of (3.8), i.e. replace one Poisson measure to another one and use the Radon-Nikodim derivative. In other words, the Poisson process under which the required characteristic of the system has a small probability, is replaced by other Poisson process under which the usual simulation becomes available, and then the required small probability is recalculated by using the Radon-Nikodim

derivative (see e.g. [22] or [18]). Recall that increments of a Poisson process are exponentially distributed, and for two exponentially distributed random variables with parameters ϑ_1 and ϑ_2, the Radon-Nikodim derivative is $\frac{\mathrm{d}\mathbb{P}_{\vartheta_1}(x)}{\mathrm{d}\mathbb{P}_{\vartheta_2}(x)} = \frac{\vartheta_1}{\vartheta_2}\mathrm{e}^{(\vartheta_2 - \vartheta_1)x}$, where $\mathbb{P}_{\vartheta_k}(x)$, $k = 1, 2$, are corresponding exponential probability distributions.

Relations (3.9) and (3.10) are derivative from (2.2) and (2.3) correspondingly.

The presence of these equations is illustrative in order to show the difference between direct simulation and simulation based on (3.8) and Theorem 4.1.

The paper is organized as follows. In Section 2 we discuss basic equations giving us then three aforementioned relations (3.8), (3.9) and (3.10) for loss proportions. In Section 3 we use semimartingale decomposition and Lenglart-Rebolledo inequality to derive then crucial relation (3.8). In Section 4 we prove Theorem 4.1 and the corollary from this theorem related to particular case where the point processes $A(t)$ and $D(t)$ are Poisson. Specifically, Theorem 4.1 establishes relationships for the proportions $\lim_{t \to \infty} t^{-1} \int_0^t \mathbb{P}\{Q_1(s) = i, Q_2(s) = j\}\mathrm{d}s$ ($i = 0, 1, \ldots, n - 1$; $j = 0, 1$). Section 5 is devoted to analysis of the performance measure based on loss proportion. There is formulated the problem of optimization of the proposed performance measure and described the algorithms for its solution. Section 6 provides two numerical examples. The concluding remarks are in Section 7.

2. Basic equations

All point processes considered in this paper are assumed to be right-continuous having the left-side limits.

The following notation is used. The arrival process is denoted $A(t)$. The queueing process (the number of occupied servers process) is denoted $Q_1(t)$. The orbit state is denoted $Q_2(t)$. It can take only the values 0 and 1 correspondingly to the cases of empty and occupied orbit space. The loss process is denoted $Q_3(t)$. Specifically, $Q_3(t)$ is the cumulated number of losses up to time t. The moments of retrials are governed by point process $D(t)$, that is, the retrial moments are $t_1, t_2, \ldots$ and $D(t) = \min\{k : t_k \leq t\}$. The sequence of independent Poisson processes all with the same rate μ is denoted $\pi_l(t)$.

Then we have the following equations ($\mathbf{I}\{\mathcal{A}\}$ denotes the indicator of event $\mathcal{A}$):

$$(2.1) \quad Q_1(t) + Q_2(t) + Q_3(t) \quad = \quad A(t) - \int_0^t \sum_{l=1}^n \mathbf{I}\{Q_1(s-) \geq l\} \mathrm{d}\pi_l(s),$$

$$(2.2) \qquad Q_2(t) + Q_3(t) \quad = \quad \int_0^t \mathbf{I}\{Q_1(s-) = n\} \mathrm{d}A(s)$$

$$- \int_0^t \mathbf{I}\{Q_1(s-) \neq n\} \mathbf{I}\{Q_2(s-) = 1\} \mathrm{d}D(s),$$

$$(2.3) \qquad Q_3(t) \quad = \quad \int_0^t \mathbf{I}\{Q_1(s-) = n\} \mathbf{I}\{Q_2(s-) = 1\} \mathrm{d}A(s).$$

Clearly, that the meaning of the left-hand side of relation (2.1) is the total number of customers in the system including customers in service, one in orbit (in the case of occupied place) as well as all losses up to time t. Notice, that the right-hand side of (2.1) is as the right-hand side of (2.1) in [1]. The first term of the right-hand side of (2.2) means the total number of arrivals to busy system where all servers are occupied. The second term of the right-hand side of (2.2) is the subtracted number of successfully retrial customers up to time t from the total number of arrivals to busy system during the same time t. Successful retrials occur in the cases where immediately before retrial instants there is at least one busy server. Notice, that relation (2.2) is structured similarly to corresponding relation (2.2) of [1]. Relation (2.3) characterizes the number of losses. Losses occur in the cases where immediately before arrival all servers are busy and the place in orbit is occupied. This is indicated by the right-hand side of (2.3).

3. Semimartingale decompositions and normalization

In this section we discuss the normalized processes by dividing these processes by t. As t increases to infinity, the behaviour of the process $Q_3(t)$ divided by t is the significant performance characteristics characterizing losses per time unit.

For the purpose of study normalized characteristics we need in semimartingale decomposition of the processes (e.g. Liptser and Shiryayev [14], Jacod and Shiryayev [11]).

The processes $A(t)$, $D(t)$ and $\pi_l(t)$, $l = 1, 2, \ldots, m$ all are assumed to be given on the common filtered probability space $(\Omega, \mathcal{F}, \mathbf{F} = (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$. All these processes are semimartingales. The semimartingale decomposition for arbitrary point process $X(t)$ is written $X(t) = \widehat{X}(t) + M_X(t)$, where $\widehat{X}(t)$ is a compensator and $M_X(t)$

is a local square integrable martingale. For example: $A(t)=\widehat{A}(t)+M_A(t)$. In some cases for local square integrable martingales we use the notation M with indexes such as $M_{i,j,k}(t)$. Such types of notation will be especially explained.

Observing first (2.1), rewrite it

$$\text{(3.1)} \qquad Q_1(t) + Q_2(t) + Q_3(t) \quad = \quad A(t) - C(t),$$

$$\text{(3.2)} \qquad C(t) \quad = \quad \int_0^t \sum_{l=1}^n \mathbf{I}\{Q_1(s-) \geq l\}\mathrm{d}\pi_l(s).$$

The semimartingale $C(t)$ admits Doob-Meyer decomposition

$$\text{(3.3)} \qquad C(t) = \widehat{C}(t) + M_C(t).$$

In turn, the compensator $\widehat{C}(t)$ admits the representation

$$\text{(3.4)} \qquad \widehat{C}(t) = \mu \int_0^t Q_1(s)\mathrm{d}s$$

(for details see Dellacherie [9], Liptser and Shiryayev [14], [15], Theorem 1.6.1). Therefore, in view of (3.2), (3.3) and (3.4), one can rewrite (3.1) as follows:

$$\text{(3.5)} \qquad Q_1(t) + Q_2(t) + Q_3(t) = A(t) - \mu \int_0^t Q_1(s)\mathrm{d}s - M_C(t).$$

Let us pass to normalized processes. For arbitrary process $X(t)$, for $t > 0$ its normalization is denoted by small letter, so

$$x(t) = \frac{X(t)}{t}.$$

Therefore, passing to normalized processes in (3.5) can we written as

$$\text{(3.6)} \qquad q_1(t) + q_2(t) + q_3(t) = a(t) - \frac{\mu}{t}\int_0^t Q_1(s)\mathrm{d}s - m_C(t).$$

Assume now that t increases to infinity. Then, because of stochastic boundedness of $Q_1(t)$ and $Q_2(t)$ the terms $q_1(t)$ and $q_2(t)$ vanish with probability 1, and the left-hand side of (3.6) in limit looks $\mathbb{P}\text{-}\lim_{t\to\infty} q_3(t)$.

Next, according to (1.2), as t increases to infinity, $a(t)$ converges with probability 1 to λ. The term $m_C(t)$ vanishes in probability, since according to Lenglart-Rebolledo inequality we have:

$$\mathbb{P}\{|m_C(t)| > \delta\} \leq \mathbb{P}\left\{\sup_{0<s\leq t}\left|\frac{sm_C(s)}{t}\right| > \delta\right\}$$

$$= \mathbb{P}\left\{\sup_{0<s\leq t}|M_C(s)| > \delta t\right\}$$

(3.7)

$$\leq \frac{\epsilon}{\delta^2} + \mathbb{P}\left\{\sum_{l=1}^{n}\pi_l(t) > \epsilon t^2\right\}$$

$$= \frac{\epsilon}{\delta^2} + \mathbb{P}\left\{\frac{1}{t}\sum_{l=1}^{n}\pi_l(t) > \epsilon t\right\},$$

and because of arbitrariness of ϵ in (3.7) the statement is right.

Therefore, applying the Lebesgue theorem on dominated convergence

$$\mathbb{P}\text{-}\lim_{t\to\infty} q_3(t) = \lambda - \mu\left(\mathbb{P}\text{-}\lim_{t\to\infty}\frac{1}{t}\int_0^t Q_1(s)\mathrm{d}s\right)$$

(3.8)

$$= \lambda - \mu\left(\lim_{t\to\infty}\frac{1}{t}\mathbb{E}\int_0^t Q_1(s)\mathrm{d}s\right)$$

The meaning of the term of the right-hand side of (3.8), taken in brackets, is the expected number of occupied servers during the long-run history, or expected stationary number of occupied servers.

Relations (2.2) and (2.3) provide additional information on the behaviour of the process $q_3(t)$. Specifically, we have

$$\mathbb{P}\text{-}\lim_{t\to\infty} q_3(t) = \lim_{t\to\infty}\frac{1}{t}\mathbb{E}\int_0^t \mathbf{I}\{Q_1(s-) = n\}\mathrm{d}A(s)$$

(3.9)

$$- \lim_{t\to\infty}\frac{1}{t}\mathbb{E}\int_0^t \mathbf{I}\{Q_1(s-) \neq n\}\mathbf{I}\{Q_2(s-) = 1\}\mathrm{d}D(s)$$

and

(3.10) $$\mathbb{P}\text{-}\lim_{t\to\infty} q_3(t) = \lim_{t\to\infty}\frac{1}{t}\mathbb{E}\int_0^t \mathbf{I}\{Q_1(s-) = n\}\mathbf{I}\{Q_2(s-) = 1\}\mathrm{d}A(s).$$

In the particular case where the process $A(t)$ is Poisson from (3.10) we obtain:

$$\mathbb{P}\text{-}\lim_{t\to\infty} q_3(t) = \lambda\lim_{t\to\infty}\frac{1}{t}\mathbb{E}\int_0^t \mathbf{I}\{Q_1(s) = n\}\mathbf{I}\{Q_2(s) = 1\}\mathrm{d}s$$

(3.11)

$$= \lambda\lim_{t\to\infty}\frac{1}{t}\int_0^t \mathbb{P}\{Q_1(s) = n, Q_2(s) = 1\}\mathrm{d}s,$$

and from (3.9) we obtain

(3.12)
$$\mathbb{P}\text{-}\lim_{t\to\infty} q_3(t) = \lambda \lim_{t\to\infty} \frac{1}{t} \int_0^t \mathbb{P}\{Q_1(s) = n\}\mathrm{d}s$$
$$- \lim_{t\to\infty} \frac{1}{t}\mathbb{E}\int_0^t \mathbf{I}\{Q_1(s-) \neq n\}\mathbf{I}\{Q_2(s-) = 1\}\mathrm{d}D(s).$$

In the case where both $A(t)$ and $D(t)$ are Poisson, $q_3(t)$ converges to its limit with probability 1 as t increases to infinity, because both $m_A(t)$ and $m_D(t)$ vanish with probability 1. Specifically, in this case we obtain:

(3.13)
$$\mathbb{P}\left\{\lim_{t\to\infty} q_3(t) = \lambda \lim_{t\to\infty} \frac{1}{t}\int_0^t \mathbb{P}\{Q_1(s) = n\}\mathrm{d}s \right.$$
$$\left. - \Delta \lim_{t\to\infty} \frac{1}{t}\int_0^t \mathbb{P}\{Q_1(s) \neq n, Q_2(s) = 1\}\mathrm{d}s\right\}$$
$$= \mathbb{P}\left\{\lim_{t\to\infty} q_3(t) = \lambda \lim_{t\to\infty} \mathbb{P}\{Q_1(t) = n\} \right.$$
$$\left. - \Delta \lim_{t\to\infty} \mathbb{P}\{Q_1(t) \neq n, Q_2(t) = 1\}\right\}$$
$$= 1.$$

There is use of the fact of existence of the limiting stationary probabilities in (3.13) as $t \to \infty$, as well as the equalities

$$\lim_{t\to\infty} \mathbb{P}\{Q_1(t) = n\} = \lim_{t\to\infty} \frac{1}{t}\int_0^t \mathbb{P}\{Q_1(s) = n\}\mathrm{d}s,$$
$$\lim_{t\to\infty} \mathbb{P}\{Q_1(t) \neq n, Q_2(t) = 1\} = \lim_{t\to\infty} \frac{1}{t}\int_0^t \mathbb{P}\{Q_1(s) \neq n, Q_2(s) = 1\}\mathrm{d}s.$$

4. LIMITING FREQUENCIES AND DISTRIBUTIONS

The results obtained in the previous section do not provide completed information about losses in this system. We have no information about the quantity $\lim_{t\to\infty} t^{-1}\int_0^t Q_1(s)\mathrm{d}s$ entering to the right-hand side of (3.8). The information about the right-hand sides of (3.9) and (3.10) is unknown as well.

Therefore in this section we derive equations for the following frequencies:

(4.1)
$$\lim_{t\to\infty} \frac{1}{t}\int_0^t \mathbb{P}\{Q_1(s) = i, Q_2(s) = j\}\mathrm{d}s,$$
$$i = 0, 1, \ldots, n; \ j = 0, 1.$$

For this purpose we introduce the processes:

$$I_{i,j}(t) = \mathbf{I}\{Q_1(t) = i \ \cap \ Q_2(t) = j\},$$

(4.2)

$$i = 0, 1, \ldots, n; \ j = 0, 1,$$

taking the value 0 if at least one of the indexes i or j is negative.

The jump of a point process is denoted by adding $\triangle$ [1]. For example, $\triangle A(t)$ is a jump of $A(t)$, $\triangle \pi_l(t)$ is a jump for $\pi_l(t)$ and so on.

Denote

$$\Pi_i(t) = \sum_{l=1}^{i} \pi_l(t).$$

Then we have the following equations:

(4.3)
$$\mathbf{I}\{Q_1(t-) + \triangle Q_1(t) = i \ \cap \ Q_2(t) = Q_2(t-) = j\}$$
$$= \ I_{i-1,j}(t-)\triangle A(t) + I_{i+1,j}(t-)\triangle\Pi_{i+1}(t)$$
$$+ \ I_{i,j}(t-)[1 - \triangle A(t)][1 - \triangle\Pi_i(t)][1 - j\triangle D(t)],$$
$$i = 0, 1, \ldots, n-1; \ j = 0, 1,$$

$$\mathbf{I}\{Q_1(t-) + \triangle Q_1(t) = i \ \cap Q_2(t-) = 1 \ \cap \ Q_2(t) = 0\}$$
(4.4)
$$= I_{i-1,1}(t-)\triangle D(t),$$
$$i = 0, 1, \ldots, n-1,$$

(4.5)
$$\mathbf{I}\{Q_1(t-) + \triangle Q_1(t) = n \ \cap \ Q_2(t) = Q_2(t-) = 0\}$$
$$= I_{n-1,0}(t-)\triangle A(t) + I_{n,0}(t-)[1 - \triangle\Pi_n(t)][1 - \triangle A(t)],$$

(4.6)
$$\mathbf{I}\{Q_1(t-) + \triangle Q_1(t) = n \ \cap Q_2(t-) = 0 \ \cap \ Q_2(t) = 1\}$$
$$= I_{n,0}(t-)\triangle A(t),$$

(4.7)
$$\mathbf{I}\{Q_1(t-) + \triangle Q_1(t) = n \ \cap \ Q_2(t) = Q_2(t-) = 1\}$$
$$= I_{n-1,1}(t-)\triangle A(t) + I_{n,1}(t-)\triangle A(t) + I_{n,1}(t-)[1 - \triangle A(t)][1 - \triangle\Pi_n],$$

---

[1]This is the triangle sign. We hope that this sign will not be mixed with the parameter (capital Greek letter) Δ

$$\mathbf{I}\{Q_1(t-) + \triangle Q_1(t) = n \ \cap \ Q_2(t-) = 1 \ \cap \ Q_2(t) = 0\}$$

(4.8)

$$= I_{n-1,1}(t-)\triangle D(t).$$

Then,

$$\triangle I_{i,j}(t) = \mathbf{I}\{Q_1(t-) + \triangle Q_1(t) = i \ \cap \ Q_2(t) = Q_2(t-) = j\}$$

$$+ \mathbf{I}\{Q_1(t-) + \triangle Q_1(t) = i \ \cap \ Q_2(t-) \neq Q_2(t) = j\}$$

(4.9)

$$- I_{i,j}(t-),$$

$$i = 0, 1, \ldots, n; \ j = 0, 1.$$

Since

$$\sum_{s \leq t} \triangle I_{i,j} = I_{i,j}(t) - I_{i,j}(0),$$

then taking into account (4.3) - (4.9) we have the following.

For $i = 0, 1, \ldots, n-1$,

$$I_{i,j}(t) = I_{i,j}(0) + \int_0^t [I_{i-1,j}(s-) - I_{i,j}(s-)]\mathrm{d}A(s)$$

(4.10)

$$- \int_0^t I_{i,j}(s-)\mathrm{d}\Pi_i(s) - j \int_0^t I_{i,j}(s-)\mathrm{d}D(s)$$

$$+ \int_0^t I_{i+1,j}(s-)\mathrm{d}\Pi_{i+1}(s) + (1-j) \int_0^t I_{i-1,1}(s-)\mathrm{d}D(s).$$

For $i = n$ we have the following pair of equations. In the case $j = 0$ we have:

$$I_{n,0}(t) = I_{n,0}(0) + \int_0^t [I_{n-1,0}(s-) - I_{n,0}(s-)]\mathrm{d}A(s)$$

(4.11)

$$- \int_0^t I_{n,0}(s-)\mathrm{d}\Pi_n(s)$$

$$+ \int_0^t I_{n-1,1}(s-)\mathrm{d}D(s).$$

In turn, in the case $j = 1$ we have:

$$I_{n,1}(t) = I_{n,1}(0) + \int_0^t I_{n-1,1}(s-)\mathrm{d}A(s)$$

(4.12)

$$+ \int_0^t I_{n,0}(s-)\mathrm{d}A(s) - \int_0^t I_{n,1}(s-)\mathrm{d}\Pi_n(s).$$

By semimartingale decomposition, from (4.10), (4.11) and (4.12) we obtain the following equations. In the cases $i = 0, 1, \ldots, n - 1$; $j = 0, 1$ we have:

$$
\begin{aligned}
I_{i,j}(t) =&\, I_{i,j}(0) + \int_0^t [I_{i-1,j}(s-) - I_{i,j}(s-)] \mathrm{d}A(s) \\
&- i\mu \int_0^t I_{i,j}(s) \mathrm{d}s + (i+1)\mu \int_0^t I_{i+1,j}(s) \mathrm{d}s \\
&- j \int_0^t I_{i,1}(s-) \mathrm{d}D(s) + (1-j) \int_0^t I_{i-1,1}(s-) \mathrm{d}D(s) \\
&+ M_{i,j}(t),
\end{aligned}
$$
(4.13)

with the martingale

$$
(4.14) \qquad M_{i,j}(t) = -\int_0^t I_{i,j}(s-) \mathrm{d}M_{\Pi_i}(s) + \int_0^t I_{i+1,j}(s-) \mathrm{d}M_{\Pi_{i+1}}(s).
$$

For $i = n$ we have the following pair of equations. In the case $j = 0$ we have:

$$
\begin{aligned}
I_{n,0}(t) =&\, I_{n,0}(0) + \int_0^t [I_{n-1,0}(s-) - I_{n,0}(s-)] \mathrm{d}A(s) \\
&- n\mu \int_0^t I_{n,0}(s) \mathrm{d}s + \int_0^t I_{n-1,1}(s-) \mathrm{d}D(s) \\
&+ M_{n,0}(t),
\end{aligned}
$$
(4.15)

with the martingale

$$
(4.16) \qquad M_{n,0}(t) = -\int_0^t I_{n,0}(s-) \mathrm{d}M_{\Pi_n}(s).
$$

In turn, in the case $j = 1$ we have:

$$
\begin{aligned}
I_{n,1}(t) =&\, I_{n,1}(0) + \int_0^t I_{n-1,1}(s-) \mathrm{d}A(s) - n\mu \int_0^t I_{n,1}(s) \mathrm{d}s \\
&+ \int_0^t I_{n,0}(s-) \mathrm{d}A(s) + M_{n,1}(t),
\end{aligned}
$$
(4.17)

with the martingale

$$
(4.18) \qquad M_{n,1}(t) = -\int_0^t I_{n,1}(s-) \mathrm{d}M_{\Pi_n}(s).
$$

Theorem 4.1. *For the limiting state frequencies of the processes $Q_1(t)$ and $Q_2(t)$ we have the following equations.*

In the boundary cases $i = 0$ for $j = 0, 1$ we have:

$$\mu \lim_{t \to \infty} \frac{1}{t} \int_0^t \mathbb{P}\{Q_1(s) = 1, Q_2(s) = j\} \, ds$$

$$(4.19) \qquad = \lim_{t \to \infty} \frac{1}{t} \mathbb{E} \int_0^t \mathbf{I}\{Q_1(s-) = 0, Q_2(s-) = j\} \, dA(s)$$

$$+ j \lim_{t \to \infty} \frac{1}{t} \mathbb{E} \int_0^t \mathbf{I}\{Q_1(s-) = 0, Q_2(s-) = j\} \, dD(s).$$

In the cases $i = 1, 2, \ldots, n-1$, $j = 0, 1$ we have:

$$i\mu \lim_{t \to \infty} \frac{1}{t} \int_0^t \mathbb{P}\{Q_1(s) = i, Q_2(s) = j\} \, ds$$

$$- (i+1)\mu \lim_{t \to \infty} \frac{1}{t} \int_0^t \mathbb{P}\{Q_1(s) = i+1, Q_2(s) = j\} \, ds$$

$$= \lim_{t \to \infty} \frac{1}{t} \mathbb{E} \int_0^t [\mathbf{I}\{Q_1(s-) = i-1, Q_2(s-) = j\}$$

$$(4.20) \qquad \qquad - \mathbf{I}\{Q_1(s-) = i, Q_2(s-) = j\}] \, dA(s)$$

$$+ (1-j) \lim_{t \to \infty} \frac{1}{t} \mathbb{E} \int_0^t \mathbf{I}\{Q_1(s-) = i-1, Q_2(s-) = 1\} \, dD(s)$$

$$- j \lim_{t \to \infty} \frac{1}{t} \mathbb{E} \int_0^t \mathbf{I}\{Q_1(s-) = i, Q_2(s-) = 1\} \, dD(s).$$

In the case $i = n$ and $j = 0$ we have:

$$n\mu \lim_{t \to \infty} \frac{1}{t} \int_0^t \mathbb{P}\{Q_1(s) = n, Q_2(s) = 0\} \, ds$$

$$= \lim_{t \to \infty} \frac{1}{t} \mathbb{E} \int_0^t [\mathbf{I}\{Q_1(s-) = n-1, Q_2(s-) = 0\}$$

$$(4.21) \qquad \qquad - \mathbf{I}\{Q_1(s-) = n, Q_2(s-) = 0\}] \, dA(s)$$

$$+ \lim_{t \to \infty} \frac{1}{t} \mathbb{E} \int_0^t \mathbf{I}\{Q_1(s-) = n, Q_2(s-) = 1\} \, dD(s).$$

In the case $i = n$ and $j = 1$ we have:

$$n\mu \lim_{t \to \infty} \frac{1}{t} \int_0^t \mathbb{P}\{Q_1(s) = n, Q_2(s) = 1\} \, ds$$

$$(4.22) \qquad = \lim_{t \to \infty} \frac{1}{t} \mathbb{E} \int_0^t \mathbf{I}\{Q_1(s-) = n-1, Q_2(s-) = 1\} \, dA(s)$$

$$+ \lim_{t \to \infty} \frac{1}{t} \mathbb{E} \int_0^t \mathbf{I}\{Q_1(s-) = n, Q_2(s-) = 0\} \, dA(s).$$

Proof. The proof of this theorem easily follows from representations (4.13), (4.15) and (4.17). We divide the left- and right-hand sides by t, and pass to the limits in probability as t increases to infinity.

Specifically, (4.19) and (4.20) follow from (4.13) and from the fact that the martingales $m_{i,j}(t)$ vanish with probability 1 as $t \to \infty$. The last is the consequence of vanishing $m_{\Pi_i}(t)$ and $m_{\Pi_{i+1}}(t)$ with probability 1 as $t \to \infty$. (The right-hand side of (4.14) is divided by t, and then we pass to the limit with probability 1 as $t \to \infty$.) Relations (4.21) and (4.22) follow similarly from the corresponding relations (4.15) and (4.17) and vanishing the martingales $m_{n,0}(t)$ and $m_{n,1}(t)$ given by (4.16) and (4.18). We also use the Lebesgue dominated convergence theorem to replace $\mathbb{P}\text{-}\lim_{t\to\infty}$ by $\lim_{t\to\infty} \mathbb{E}$ in the places where it is required. $\qquad\square$

The standard particular cases from Theorem 4.1 are the cases where $A(t)$ is Poisson, $D(t)$ is Poisson, and both $A(t)$ and $D(t)$ are Poisson all together. We do not consider the first two cases, and the only last of these three cases is formulated and proved below. In this case the representation is deduced explicitly and has the form of limiting distributions rather than frequencies.

Corollary 4.2. *In the case where $A(t)$ is Poisson with rate λ, and $D(t)$ is Poisson with rate Δ denote $P_{i,j} = \lim_{t\to\infty} \mathbb{P}\{Q_1(t) = i, Q_2(t) = j\}$. We have the following.*

In the boundary cases $i = 0$ for $j = 0, 1$ we have:

$$(4.23) \qquad\qquad \mu P_{1,j} = \lambda P_{0,j} + j\Delta P_{0,j}.$$

In the cases $i = 1, 2, \ldots, n-1$, $j = 0, 1$ we have:

$$(4.24) \qquad i\mu P_{i,j} - (i+1)\mu P_{i+1,j} = \lambda P_{i-1,j} - \lambda P_{i,j} + (1-j)\Delta P_{i-1,1} - jP_{i,1}.$$

In the case $i = n$ and $j = 0$ we have:

$$(4.25) \qquad\qquad n\mu P_{n,0} = \lambda P_{n-1,0} - \lambda P_{n,0} + \Delta P_{n,1}.$$

In the case $i = n$ and $j = 1$ we have:

$$(4.26) \qquad\qquad n\mu P_{n,1} = \lambda P_{n-1,1} + \lambda P_{n,0}.$$

Proof. In the case where $A(t)$ and $D(t)$ are Poisson, the proof of this corollary is based on the following two relations:

$$(4.27) \qquad \lim_{t\to\infty} \frac{1}{t}\mathbb{E}\int_0^t \mathbf{I}\{Q_1(s-) = i, Q_2(s-) = j\}\mathrm{d}A(s) = \lambda P_{i,j},$$
$$i = 0, 1, \ldots, n;\ \ j = 0, 1,$$

and

$$\lim_{t\to\infty} \frac{1}{t}\mathbb{E} \int_0^t \mathbf{I}\{Q_1(s-) = i, Q_2(s-) = j\}\mathrm{d}D(s) = \Delta P_{i,j},$$

(4.28)

$$i = 0, 1, \ldots, n; \ j = 0, 1.$$

Then inserting (4.27) and (4.28) into (4.19)-(4.22) leads to the statement of the corollary. Thus there is only required to prove (4.27) and (4.28).

Prove (4.27). By semimartingale decomposition of Poisson process we have $A(t) = \lambda t + M_A(t)$. Substituting this representation for (4.27) we obtain:

(4.29)
$$\lim_{t\to\infty} \frac{1}{t}\mathbb{E} \int_0^t \mathbf{I}\{Q_1(s-) = i, Q_2(s-) = j\}\mathrm{d}A(s)$$
$$= \lim_{t\to\infty} \frac{\lambda}{t} \int_0^t \mathbb{P}\{Q_1(s-) = i, Q_2(s-) = j\}\mathrm{d}s$$
$$+ \lim_{t\to\infty} \frac{1}{t}\mathbb{E} \int_0^t \mathbf{I}\{Q_1(s-) = i, Q_2(s-) = j\}\mathrm{d}M_A(s)$$

$$i = 0, 1, \ldots, n; \ j = 0, 1.$$

The first term of the right-hand side is equal to $\lambda P_{i,j}$, and the second term is equal to 0. Therefore (4.27) is proved.

The proof of (4.28) is analogous. The corollary is proved. $\square$

5. Performance analysis and algorithms

One of the most important performance characteristics is the loss proportion per arriving customer. Specifically, denoting the loss proportion by f, we write

(5.1)
$$f = \frac{\mathbb{P}\text{-}\lim_{t\to\infty} q_3(t)}{\mathbb{P}\text{-}\lim_{t\to\infty} a(t)} = \frac{1}{\lambda} \ \mathbb{P}\text{-}\lim_{t\to\infty} q_3(t).$$

The parameters λ, μ and Δ are assumed to be given. So, the problem is to find the appropriate number of servers n such that $f \leq \alpha$. More accurately, the problem is to minimize n subject to $f \leq \alpha$.

In general case, we have no explicit representation for the processes, and then simulation techniques are used. There are three equivalent relations obtained for $\mathbb{P}\text{-}\lim_{t\to\infty} q_3(t)$ in Section 2: (3.8), (3.9) and (3.10).

The simplest of them looks (3.10). According to (3.10) the appropriate limit $\mathbb{P}\text{-}\lim_{t\to\infty} q_3(t)$ can be obtained straightforwardly. Immediately before each arrival (one occurring say at instant s) we only check the event $\{Q_1(s-) = n \cap Q_2(s-) = 1\}$. Such type of simulation need not require any theory. However, in the case of

small α this simulation is not effective. As the large number of servers is chosen, and the loss proportion should be small, resulting in erroneous conclusion.

The other relation is given by (3.9). For a small α it is also based on the rare events, and its application is impossible.

In this general case only (3.8) is available. It is based on the average number of busy servers $\lim_{t\to\infty} t^{-1}\mathbb{E}\int_0^t Q_1(s)\mathrm{d}s$ which can be estimated by simulation. More specifically, the simulation is based on application of Theorem 4.1, helping us to obtain the frequencies

$$\lim_{t\to\infty}\frac{1}{t}\int_0^t \mathbb{P}\{Q_1(s)=i\}\mathrm{d}s = \lim_{t\to\infty}\frac{1}{t}\int_0^t \mathbb{P}\{Q_1(s)=i, Q_2(s)=0\}\mathrm{d}s$$

(5.2)
$$+ \lim_{t\to\infty}\frac{1}{t}\int_0^t \mathbb{P}\{Q_1(s)=i, Q_2(s)=1\}\mathrm{d}s,$$

$$i = 0, 1, \ldots, n.$$

The first and second terms of the right-hand size are obtained from Theorem 4.1. They are expressed via the frequencies

$$\lim_{t\to\infty}\frac{1}{t}\mathbb{E}\int_0^t \mathbf{I}\{Q_1(s-)=i, Q_2(s-)=j\}\mathrm{d}A(s)$$

$$\lim_{t\to\infty}\frac{1}{t}\mathbb{E}\int_0^t \mathbf{I}\{Q_1(s-)=i, Q_2(s-)=j\}\mathrm{d}D(s)$$

$$i = 0, 1, \ldots, n; \; j = 0, 1,$$

which in turn are easily evaluated by simulation.

It is worth noting that direct simulating of $\lim_{t\to\infty} t^{-1}\int_0^t \mathbb{P}\{Q_1(s)=i, Q_2(s)=j\}\mathrm{d}s$ $(i = 0, 1, \ldots, n, \; j = 0, 1)$ and consequently $\lim_{t\to\infty} t^{-1}\mathbb{E}\int_0^t Q_1(s)\mathrm{d}s$ is not available. (See Section 9 of [1] for the detailed discussion of this question.)

Thus, for the simulation purposes we use (3.8) and the relations of Theorem 4.1.

Recall that the problem is to minimize m subject to $f \le \alpha$. The upper bound for n is unknown, and the special search procedure is necessary. The relevant search procedure is known due to Rubalskii [21], who proposed the search algorithm for minimization of a unimodal function on an unbounded set. The optimal algorithm is an extension of the standard Fibonacci procedure.

Thus, solution of the problem is based on two main steps:

• Simulation

• Search step

These procedures are repeated until the optimal solution is not found.

In the particular case where both $A(t)$ and $D(t)$ are Poisson, for stationary probabilities $P_{i,j}$, $i = 0, 1, \ldots, n$; $j = 0, 1$, we have the system of algebraic equations. Furthermore, the upper bound for the value n can be evaluated as well. In order to evaluate the upper bound for n, one can imagine that arrival rate is $\lambda + \Delta$, that is retrials occur all the time continuously and the retrial space is always occupied. In this case the loss probability is greater then that original, and consequently the corresponding value n is greater than that required. This upper bound n can be easily calculated by the Erlang loss formula. Then, having the upper and lower bounds we obtain the optimal number of servers easier.

The similar approach can be used and for non-Markovian system. In the most practical cases where input stream is recurrent the upper bound for n can be easily evaluated by known formulae for the loss systems.

However in general we have to use one or other special algorithm, say the method of [21], in order to find optimal value of n.

6. Numerical examples

In this section we provide numerical solution for two systems. The first system is the Markovian system with $\lambda = 10$, $\Delta = 2$, $\mu = 1$ and $\alpha = 0.0001$ In the second example we assume that the increments of the processes $A(t)$ and $D(t)$ are deterministic with the same parameters.

6.1. **Example 1.** Assuming that the lower bound for m is equal to 1, let us find the upper bound. For this purpose we consider the M/M/n/0 queueing system with input rate $12 = \lambda + \Delta$ and $\mu = 1$. According to Erlang loss formula, for the loss probability we have

$$p_n = \frac{\dfrac{\rho^n}{n!}}{\sum_{i=0}^n \dfrac{\rho^i}{i!}},$$

where $\rho = \lambda/\mu = 12$. Assuming that the loss probability is equal to $\alpha = 0.0001$ let us find the value n.

For $n=26$ the loss probability is 0.000174, and for $n=27$ the loss probability is 0.000078 The first of these probabilities is greater than 0.0001, and the second one is smaller than 0.0001 Therefore, the upper bound for n is 27.

Let us now find n optimal. The solution of this problem can be achieved as follows. We find a new value of $n=1+(27-1)/2=14$. Now we solve the system of equations (4.23)-(4.26) given by Corollary 4.2. We also use (5.1) and (3.13). For loss probability we obtain 0.064841 Therefore in the next step we find a new value n by $14+(27-14)/2=20.5$ taking then the integer part. With new value $n = 20$ we solve the system again, and for the loss probability we obtain 0.001226 This value is greater than α, and therefore the new value of n is the integer part of $20+(27-20)/2$ equal to 23. With this value of n for the loss probability we have 0.000110 Thus the last step is anticipated with $n=24$. (Formally we must analyze first the value $n=25$ and only then accept $n=24$.) For this value n the loss probability is 0.000045 Thus the problem is solved, and the value $n=24$ is the optimal number of servers decreasing the loss probability to the required value.

6.2. **Example 2.** Assume that all parameters are the same as in Example 1, but increments of the processes $A(t)$ and $D(t)$ are deterministic. In this case the scheme of calculation includes simulation as it is explained in Section 5. For this concrete system the algorithm can be simplified. Specifically, as in Example 1 we first evaluate the upper bound for n. For our approximation of the upper bound of n we consider the D/M/n/0 queueing system, the interarrival times of which all are equal to $1/(\lambda + \Delta)$, i.e. in our case 1/12. Notice, that under the assumption that customers constantly arrive to the main queue by $A(t)$ and $D(t)$ we do not have the model of D/M/n/0 queue exactly. Although input stream is based on two sources with deterministic interarrival times, the structure of the overall input stream is complicated. The arguments for such approximating by the D/M/n/0 queueing system are heuristic. Nevertheless, such approximation is available. This is supported by computations as well.

The loss probability for GI/M/n/0 loss systems is well-known (e.g. [8], [19], [20], [24] as well as [5]):

$$(6.1) \qquad p_n = \left[\sum_{i=0}^{n} \binom{n}{i} \prod_{j=1}^{i} \frac{1 - r_j}{r_j} \right]^{-1},$$

where

$$r_j = \int_0^\infty \mathrm{e}^{-j\mu x} \mathrm{d}R(x),$$

and $R(x)$ is the probability distribution function of an interarrival time. In our case interarrival times are deterministic of length 1/12 and $\mu = 1$. Therefore, in (6.1)

$$r_j = \mathrm{e}^{-j/12}.$$

According to our calculations the upper bound for n is 22. The value of the loss probability is 0.000036 Let us note that for $n=21$ the value of the loss probability is 0.000137, i.e. it is slightly greater than $\alpha=0.0001$

In the next step we choose then the integer part of $1+(22\text{-}1)/2$, and the value n is 11. Simulating with this value of n yields the value of the loss probability 0.125444 Therefore in the next step we choose the integer part of $11+(22\text{-}11)/2$, i.e. $n=16$. With this value of n by simulating we have the value of loss probability 0.002784 In the next step $n=16+(22\text{-}16)/2=19$. With this value n we correspondingly obtain the loss probability 0.000060 In the next step we choose the integer part of 19-(19-16)/2 for n, i.e. $n=17$. The loss probability in this case is 0.000786 There is one value $n=18$ which must be checked. For this value the loss probability is 0.000214 Thus our conclusion is $n=17$. Recall that the loss probability in this case is 0.000060

7. Concluding remarks

In this paper we studied multiserver retrial queueing system having only one space in orbit. For this queueing system we derived analytic results permitting us to provide performance analysis. Specifically we solved the problem to find the possibly minimal number of servers decreasing the loss probability to given small value. In general case the algorithm of solution requires effective simulation, meeting the challenge of rare events problem. Numerical examples were provided for two cases. In the first case all processes were assumed to be Poisson, and the system Markovian. For Markovian system the calculations were based on analytic results. In the second case the processes $A(t)$ and $D(t)$ had deterministic increments. The analysis in this case was based on effective simulation. Comparison of these results showed that in the case of Markovian system the required number of servers, making the loss proportions smaller than the given value, is relatively greater than that in the system having the processes $A(t)$ and $D(t)$ with deterministic increments.

References

[1] ABRAMOV, V.M. (2006). Analysis of multiserver retrial queueing system: A martingale approach and an algorithm of solution. *Annals of Operations Research*, 141, 19-50.

[2] ARTALEJO, J.R. (1999). Accessible bibliography on retrial queues. *Mathematical and Computer Modelling*, 30, 1-6.

[3] ARTALEJO, J.R. AND FALIN, G.I. (2002). Standard and retrial queueing systems. A comparative analysis. *Revista Matematica Complutense*, 15, 101-129.

[4] ATENCIA, I. AND PHONG, N.M. (2004). A queueing system under LCFS PR discipline with Markovian arrival process and general times of searching for service. *Revista Investigacion Operational* 25, 293-298.

[5] BHARUCHA-REID, A.T. (1960). *Elements of the Theory of Markov Processes and Their Application*. McGraw-Hill, New York.

[6] BOCHAROV, P.P., D'APICE, C., MANZO, R. AND PHONG, N.H. (2001). On retrial single-server queueing system with finite buffer and multivariate Poisson flow. *Problems of Information Transmission*, 37, 397-406.

[7] BOCHAROV, P.P., PHONG, N.M. AND ATENCIA, I. (2001). Retrial queueing systems with several input flows. *Revista Investigacion Operational*, 22, 135-143.

[8] COHEN, J.W. (1957). The full availability group of trunks with an arbitrary distribution of interarrival times and negative exponential holding time distribution. *Simon Stevin* 31, 169-181.

[9] DELLACHERIE, C. (1972). *Capacites et Processus Stochastiques*. Springer-Verlag, Berlin.

[10] FALIN, G.I. (1990). A survey on retrial queues. *Queueing Systems*, 7, 127-168.

[11] JACOD, J. AND SHIRYAYEV, A.N. (1987). *Limit Theorems for Stochastic Processes*. Springer-Verlag, Berlin.

[12] HEIDELBERGER, P. (1995). Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modelling and Computer Simulation*, 5, 43-85.

[13] KOVALENKO, I.N. (2002). The loss probability in $M/G/m$ queueing systems with T-retrials in light traffic. *Dopovidi NAN Ukrainy* (Ukrainian Academy of Sciences), ser. A, No.5, 77-80.

[14] LIPTSER, R.S. AND SHIRYAYEV, A.N. (1977/1978). *Statistics of Random Processes*. Vols I, II. Springer-Verlag, Berlin.

[15] LIPTSER, R.S. AND SHIRYAYEV, A.N. (1989). *Theory of Martingales*. Kluwer, Dordrecht.

[16] MANDELBAUM, A., MASSEY, W.A. AND REIMAN, M.I. (1998). Strong approximation for Markovian service network. *Queueing Systems*, 30, 149-201.

[17] MANDELBAUM, A., MASSEY, W.A., REIMAN, M.I., STOLYAR, A. AND RIDER, B. (2002). Queue-lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommunication Systems*, 21, 149-171.

[18] MELAMED, B. AND RUBINSTEIN, R. Y. (1998). *Modern Simulation and Modelling.* John Wiley, Chichester.

[19] PALM, C. (1943). Intensitätschwankungen im Fernsprechverkehr. *Ericsson Technics* 44, 1-189.

[20] POLLACZEK, F. (1953). Generalisation de la theorie probabiliste des systemes telephoniques sans dispositif d'attente. *Comptes Rendus de l'Academie des Sciences* (Paris), 236, 1469-1470.

[21] RUBALSKII, G.B. (1982). The search of an extremum of unimodal function of one variable in an unbounded set. *U.S.S.R. Comput. Maths. Math. Phys.* 22 (1), 8-15. Transl. from Russian: *Zhurnal Vychislitelnoi Matematiki i Matematicheskoi Fiziki* 22 (1) 10-16, 251.

[22] RUBINSTEIN, R.Y. AND SHAPIRO, A. (1993). *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method.* John Wiley, Chichester.

[23] SHAHABUDDIN, P. (1995). Rare event simulation in stochastic models. *Proceedings of the 1995 Winter Simulation Conference*, C.Alexopoulos, K.Kang, W.R.Lilegdon and D.Goldsman eds, pp.178-185, IEEE press.

[24] TAKÁCS, L. (1957). On a probability problem concerning telephone traffic. *Acta Mathematika Academia Scientiarum Hungaricae* 8, 319-324.

SCHOOL OF MATHEMATICAL SCIENCES, MONASH UNIVERSITY, BUILDING 28M, CLAYTON, VICTORIA 3800, AUSTRALIA

E-mail address: `vyacheslav.abramov@sci.monash.edu.au`